Université Joseph Fourier
PolyTech Grenoble

Année Universitaire 2003-2004
RICM 3
Systèmes d'Information Décisionnels
Examen

Concepteurs : Didier DONSEZ
Date : Mars 2004
Durée : 2 heures
Remarques : Calculatrice **RECOMMANDEE** et tout document autorisé
Conseil : Lire le sujet jusqu'au bout.
Annexe : Résultats du benchmark TPC/H

Modélisation décisionnelle des actions de téléspectateurs interactifs

+ CORRECTION

L'opérateur BreeFox[1] propose à ses abonnés un boiter routeur ADSL qui groupe un accès à IP, le téléphone et un bouquet de chaînes de télévision interactives. Nous ne nous intéresserons qu'à cette dernière fonction dans notre entrepôt de données.

Les chaînes de télévision proposées par BreeFox dans son bouquet sont des chaînes nationales et des chaînes à péage. Seulement, les interruptions publicitaires (des chaînes nationales et à péage) peuvent être personnalisées en fonction du profil du foyer (ou des adultes du foyer quand le contrôle parental[2] est déverrouillé). Le téléspectateur peut réagir (interactivement) de plusieurs manières aux émissions et aux publicités qu'il reçoit au moyen de sa télécommande (ie *zapette)*:

- A tout moment, il peut zapper vers une autre chaîne
- A tout moment, il peut éteindre son poste
- Pendant une publicité, il peut zapper à la pub suivante sans attendre la fin de la publicité en cours (la durée de visualisation est importante).
- Pendant une publicité, il peut demander plus d'information sur le produit présenté (la durée de visualisation est importante) puis revenir à l'émission en cours .

[1] Toute ressemblance avec une société existante est purement fortuite !
[2] Pour éviter que les enfants du foyer apprennent trop jeunes la Biologie !

L'entrepôt de données est centré sur les actions du téléspectateur. L'objectif de cet entrepôt est de profiler au plus précis le foyer (ou plutôt le téléspectateur qui détient la télécommande) afin de maximiser la demande d'information sur les produits présentés par les publicités. En effet, l'opérateur perçoit plus d'argent de la part de l'annonceur quand le spectateur demande plus d'information au moment au l'annonce est passée !
Le schéma de l'entrepôt est constitué des tables suivantes (les clés primaires sont soulignées)

Date(CléDate, Année, Mois, JourDeMois, JourDeSemaine, TrancheHoraire, Heure, Minute, DrapeauVacances, Drapeau Manisfestation)
Foyer(CléFoyer, NomAbonné, AnnéeNaissanceAbonné, Région, Département, District, Ville, Quartier, SituationFamille, RevenuFoyer, CatégorieSocioProfessionnel, SousCatégorieSocioProfessionnel, DomaineActivité, NombreAdulte, NombreEnfant)
Emission(CléEmission, Chaîne, DateDébut, DuréeSeconde, TypeEmission, Catégorie, Annonceur)
TypeEmission= "Programme", "AnnoncePublicitaire"
Catégorie= "Météo", "Journal", "Variété", "Jeu", "JeuAvecPognion", "Film", "Foot", …
Annonceur=seulement pour les annonces publicitaires
Action(CléDate, CléFoyer, CléEmission, TypeAction, DuréeAction, DuréeRestante, DrapeauCtrlParentalDéverrouillé)

- TypeAction= « Zap Autre Chaîne », « Eteindre », « Allumer », « Zap Pub Suivante », « Demande Information », « Passif » (ie pas d'action sur la zappette)
- **DuréeAction** = Durée écoulée entre le début de l'émission et le début de l'action. ***Quand*** *TypeAction est différent de « Demande Information »*
- **DuréeAction** = Durée écoulée entre le début de l'émission et la fin du temps supplémentaire d'information avant que le spectateur regarde avant de retourner à son émission, ***Quand*** *TypeAction est égal de « Demande Information »*
- DuréeRestante = Durée restante entre l'action et la fin de l'émission (=0 si TypeAction= « Passif »
- DrapeauCtrlParentalDéverrouillé indique le contrôle parental a été déverrouillé

Rétro-Conception

Q1: Quelle est la table de fait dans cet entrepôt ?. Justifiez en 2 lignes !
Action (car au centre des dimensions, attributs additifs ou numériques)

Q2: Que pensez vous de l'attribut TypeAction de Action ?
N'est pas une mesure ! C'est une dimension dégénérée

Q3: A votre avis, il y a t'il des dimensions douteuses dans cet entrepôt ? Rappelez la définition et justifiez en 3 lignes
Foyer !.

Q4: Donnez les nouvelles tables si on décide de diminuer la taille de la table Foyer par une mini-dimension démographique

On crée une table de dimension Demographie avec les attributs de Foyer (Région, Département, District, Ville, Quartier, RevenuAssuré, RevenuFoyer, CatégorieSocioProfessionnel, SousCatégorieSocioProfessionnel, DomaineActivité ou d'autres éventuellement)
On supprime ces attributs de Foyer
On ajoute une clé de mini-dimension à la table de fait et à la table de dimension Foyer

Dimensionnement

Q5: Donnez le nombre de faits présents dans la table de fait.

Nombre de foyers abonnés	3 Millions
Nombre de actions par foyer et par heure	10
Un foyer regarde la télévision 320 jours par an, 5 heures par jour	
Nombre de tranche horaire	24
Nombre d'années	3
Taille des clés	4 octets
Taille des attributs numériques	4 octets
Taille des attributs discrets (comme les types !)	1 octet
Taille des attributs booléens (comme les drapeaux !)	1 octet

Nombre d'actes=3.000.000*10*320*5*3= 144 000 000 000 actions ou enregistrements

Donnez la taille d'un enregistrement de la table de fait ?

(4*3 clés + 4*2 attributs numériques + 1*1 attributs type + 1*1 attributs booléens)=22 octets par fait

Donnez la taille (en Octets) de stockage de la table de fait.

Taille de la table de fait= 3168000000000octets soit 3 To

Configuration Matérielle

Q6: A partir des résultats du benchmark TPC/H (http://www.tpc.org/tpch/results/tpch_results.xls) donné en annexe, choisissez la configuration matérielle et logicielle (complète) qui est la plus adaptée à votre infocentre pour une performance minimale de 25000 QphH ? Quels sont vos critères de choix ?

Remarque : vous négligerez la taille des tables de dimensions.

On choisira un SF=3000 (3000Go)
HP HP Integrity Superdome Enterprise Server 2 3000 45248 109 4922070 US $ Oracle Database 10g Enterprise Edition HP UX 11.i 64-bit

Rapports

Q7: Donnez la requête SQL qui donne le temps cumulé pour chaque type d'émission et pour chaque tranche horaire

```
SELECT D.TrancheHoraire, E.TypeEmission, SUM(DureeAction) AS TempsCumulé
FROM Emission E JOIN Action A USING (CléEmission) JOIN Date D USING (CléDate)
GROUP BY D.TrancheHoraire, E.TypeEmission
SORT BY D.TrancheHoraire ASC, TempsCumulé DESC
```

Q8: Donnez la requête SQL qui donne le top 10 des type d'émission les plus regardées en temps cumulé

```
SELECT E.TypeEmission, SUM(DureeAction) AS TempsCumulé
FROM Emission E JOIN Action A USING (CléEmission)
GROUP BY E.TypeEmission
SORT BY TempsCumulé DESC
TOP(10)
```

Q9: Donnez le rapport mensuel de 'progression' du nombre de demande d'information (ie TypeAction =« Demande Information ») et de la durée associée regardée

```
SELECT D.Année, D.Mois,
  COUNT(*) AS NombreTotal,
  SUM(A.DuréeAction) AS DuréeTotale
FROM Action A JOIN Date D USING (CléDate)
WHERE A.TypeAction=« Demande Information »
GROUP BY D.Année, D.Mois
SORT BY D.Année ASC, D.Mois ASC
```

Q10: Donnez le rapport précédent mais avec une moyenne glissante avec les 2 mois précédents.

```
SELECT R.Année, R.Mois,
        AVG(R.NombreTotal) OVER (
                ORDER BY D.Année D.Mois ASC
                ROWS 2 PRECEDING) AS MoyGlissanteNombreTotal,
        AVG(R.DuréeTotale) OVER (
                ORDER BY D.Année D.Mois ASC
                ROWS 2 PRECEDING) AS MoyGlissante DuréeTotale
FROM(
SELECT D.Année, D.Mois,
        COUNT(*) AS NombreTotal,
        SUM(A.DuréeAction) AS DuréeTotale
  FROM Action A JOIN Date D USING (CléDate)
  WHERE A.TypeAction=« Demande Information »
  GROUP BY D.Année, D.Mois
  SORT BY D.Année ASC, D.Mois ASC
) R
GROUP BY R.Année, R.Mois
SORT BY R.Année ASC, R.Mois ASC
```

Conception physique

Q11 : Rappelez le principe d'index B-Tree et celui d'un index BitMap ? Lequel est il généralement mieux adapté aux entrepôts de données ?(en 10 lignes)

Q12 : Pourquoi est il intéressant d'avoir des enregistrements de taille fixe pour la table de fait (justifiez en 5 lignes)

On peut y appliquer des indexs bitmaps !

Q13 : Est il judicieux d'utiliser les indexs bitmap sur les fichiers (stockage physique des tables) dont les enregistrements ont un taille variable (justifiez en 5 lignes)

Non, car pas d'accès aléatoire possible.

Q14 : Pourquoi est il intéressant de partitionner la table de fait sur des disques différents ? (2 raisons à justifier en 8 lignes)

1. pour les perfs
2. pour la business logique de duree de vie (validite) des donnees. Permet de purger les données anciennes (hors du champs des études)